

Description

METHOD FOR EFFICIENTLY CHECKING COVERAGE OF RULES DERIVED FROM A LOGICAL THEORY

BACKGROUND OF INVENTION

- [0001] The invention relates to a method for learning systems, and in particular, to learning systems that generate rules by derivation from a logical theory.
- [0002] Several methods for rule induction are known in the art. Common for these methods is that a set of logical rules are generated from a set of examples, where each example has been given a label, which can either be a categorical value or a numeric value. Each logical rule consists of a condition part that in turn consists of a set of logical tests as well as a conclusion part, which assigns a value for the label. For example, the condition part may be that the number of atoms must exceed the numerical value five and the molecule weight must not be less than 100 to generate a positive class. The examples may include at-

tributes, such as molecule weight, that correspond to the condition part of the logical rule. One of the most common techniques for rule induction is known as decision tree induction, which generates a set of hierarchically organized rules, where none of the rules overlap (i.e., the conditions of two different rules are mutually exclusive).

Examples of such techniques are ID3 and CART. Other techniques, such as covering or separate-and-conquer, may generate overlapping rules. Examples of such techniques are CN2 and RIPPER.

- [0003] Most techniques for rule induction allow examples to be represented as fixed-length attribute-value vectors, and the conditions to consist of simple tests that, for example, checks whether a particular attribute has a particular value, or whether the value is below or above a particular threshold. Some techniques also allow examples to be represented by arbitrary logical terms, including lists and trees, and conditions to consist of arbitrary logical literals such as tests that involve an arbitrary number of variables using arbitrarily defined predicates. Such techniques are studied primarily in a research field known as inductive logic programming and examples of such techniques are FOIL and PROGOL.

[0004] One method for rule induction is to use a logical theory from which rules are derived by using an inference procedure known as resolution. During the generation of rules to be included in the final hypothesis, a large number of candidate rules are evaluated, which involves checking for a set of training examples, which of these fulfill the conditions of the candidate rules. After the final hypothesis has been generated, it is usually applied to examples not included in the set of training examples, which again involves checking whether or not the conditions of the rules are fulfilled for each example. This is a very cumbersome process, in which complex proof trees may have to be generated repeatedly for each example. In both cases, minimizing the amount of time required to perform these tests can be of high importance. There is a need for a more effective process that does not require the repeated generation of proof trees for the examples.

[0005] The present invention provides a solution to this problem and is a method and an apparatus for efficiently checking whether or not the conditions of a rule derived by resolution from a logical theory are fulfilled by an example. The apparatus consists of the following three modules:

[0006] i) A module for generating a database from proof trees

that have been constructed from the examples using the logical theory;

- [0007] ii) A module for generating database queries from rules that have been derived from the logical theory; and
- [0008] iii) A module for querying the database with the queries obtained from the rules.

[0009] More particularly, the method is used in a computer and includes the steps of providing a logical theory that has clauses. A rule is generated that is a resolvent of clauses in the logical theory. An example is retrieved. A proof tree is generated from the example using the logical theory. The proof tree is transformed into a database of a coverage check apparatus. The rule is converted into a partial proof tree that has nodes. The partial proof tree is transformed into a database query of the coverage check apparatus. The query is executed to identify tuples in the database that correspond to the nodes of the partial proof tree. In this way, the database of pre-existing examples may be investigated to determine if a rule covers a pre-existing example, that are associated with the same logical theory, so there is no need to recreate complicated proof trees for the examples.

BRIEF DESCRIPTION OF DRAWINGS

- [0010] Fig. 1 is an overview schematic flowchart of the method of the present invention;
- [0011] Fig. 2 is a rule sequence of a logical theory of the method of the present invention;
- [0012] Fig. 3 is a schematic flowchart of a proof tree of the method of the present invention;
- [0013] Fig. 4 is a sequence for transforming proof trees into database tables according to the method of the present invention;
- [0014] Fig. 5 is a group of database tables generated from a proof tree according to the method of the present invention;
- [0015] Fig. 6 is a schematic flowchart of a derived rule according to the method of the present invention;
- [0016] Fig. 7 is a sequence for transforming a rule into a database query according to the method of the present invention;
- [0017] Fig. 8 is a database query generated according to the method of the present invention;
- [0018] Fig. 9 is a group of database tables generated from a single proof tree according to the method of the present invention;
- [0019] Fig. 10 is a database query generated assuming a single

proof tree according to the method of the present invention;

- [0020] Fig. 11 is a logical theory of the method of the present invention;
- [0021] Fig. 12 is a rule according to the method of the present invention;
- [0022] Fig. 13 is a schematic flowchart of a proof tree of the method of the present invention;
- [0023] Fig. 14 is a group of database tables generated from a proof tree according to the method of the present invention; and
- [0024] Fig. 15 is a database query generated from a rule according to the method of the present invention.

DETAILED DESCRIPTION

- [0025] With reference to Figs. 1–15 and the description below, the method of the present invention is described with reference to particular embodiments of logical theories and database systems used in connection with a computer. The present invention, however, is not limited to any particular syntax for logical theories and types of database system, nor limited by the examples described herein. Therefore, the description of the embodiments that follows is for purposes of illustration and not as a limitation.

[0026] One important feature of the method of the present invention is that the database of pre-existing examples is investigated to determine if a rule covers a pre-existing example, that is associated with the same logical theory, so there is no need to recreate complicated proof trees for the examples and the rules. Fig. 1 provides an overview of a coverage check procedure 10 of the apparatus of the present invention. The procedure has a coverage check apparatus 28 that takes as input a set of rules 14 that are resolvents of clauses 29, such as Horn clauses, in a logical theory 12 and a set of proof trees 18 that have been generated from pre-existing examples 16 that consists of atoms using the logical theory 12. In other words, the theory 12 may be used to describe which possible rules that may be created. The logical theory may function as a type of a grammar for the rules.

[0027] The procedure 10 may be used to investigate for each rule 14 and each example 16 whether or not the rule 14 covers the example 16. It then investigates whether a condition part of the rule 14 is satisfied by the example 16. The method/procedure 10 of the present invention does this by transforming the proof trees 18 into database tables of a database generator 20, by transforming the rules 14

into database queries of a query generator 22 and finally by checking in a query checker 24 whether or not the queries 22 produce empty result sets with respect to the database tables 20 to finally produce an answer 32 as the output of the apparatus 28. If a matching pre-existing example is found that is covered by the rule, there is no need to re-create the proof trees. The method and modules of the apparatus are described below.

- [0028] The method may be simplified in case each example has at most one proof tree. Finally, the procedure can be extended to handle the situation when conditions of the generated rules are not only derived from the logical theory but also contain terms derived from the examples, as indicated by a dotted line 26 in Fig. 1.
- [0029] When transforming proof trees 18 into the database tables 20 of the apparatus 28, it is assumed that all possible proof trees 18 for each example 16 have been generated using the given logical theory 12, and that each example 16 and proof tree 18 has been given a unique label.
- [0030] Fig. 2 gives an example of a logical theory 30 using a standard prolog syntax that concerns the domain of playing cards. The theory 30 may describe the pre-existing examples with the help of rules that may include hierar-

chical relations. The theory 30 also specifies which attributes are included in the pre-existing examples. A clause c1 defines a predicate reward 33 that has two arguments i.e. Color with the associated body literal color 34 and Value with the associated body literal value 36. A specific example is covered by a logical theory if it is an instance of a defined predicate, and if all corresponding body literals are covered by the logical theory, as described below. The rule or clause c1 states that something that has the predicate name reward and two arguments is covered by the logical theory, if the corresponding instances of the predicates color and value are covered by the logical theory. A clause c2 states that red is a color and clause c3 states that black is a color. A clause c4 states that face is a value and clause c5 states that numbered is a value. Clauses c6 and c7 state that hearts and diamonds are a color that is red. In other words, hearts or diamonds has a color that is red. Clauses c8 and c9 state that spades and clubs are a color that is black. In other words, spades or clubs c9 has a color that is black. Clauses c10, c11, c12 state that king, queen and knight are values that are faces. Clauses c13, .., and c22 state that 1,.. and 10 are values that are numbers.

[0031] Fig. 3 shows that logical theory 30 covers an example 38, such as reward (hearts, king). A proof tree 40, also labeled t1, can be derived from the logical theory 30 and the example 38, also labeled e1. The proof tree 40 has a first base node 43 and nodes 44, 46 in a first leg 48 and nodes, 50, 52 in a second leg 54. The proof tree may show how a specific example is covered by the logical theory. In other words, the proof tree 40 shows that example 38 is covered by the logical theory 30. More particularly, Fig. 3 shows the only possible proof tree 40 for the example 38, given the logical theory 30. It should be noted that the proof trees do not include proofs of predicates whose definitions are built into the system for generating proof trees, such as $=/2$ in the example.

[0032] Fig. 4 is a sequence 41 that has an example labeled e, a proof tree T, a proof tree labeled t and a set of database tables D as input. An output with the updated database tables D that includes tuples containing the examples e and tree t is generated. The set of database tables, which initially is an empty set, may be updated using the process sequence shown in Fig. 4. The sequence is called once for each proof tree that has been generated, and the input to each call is, besides the proof tree and its label,

the example label and the database tables generated by preceding calls to the sequence. In this way, the information of the proof trees is stored in database tables.

[0033] Fig. 5 shows a table group 42 including the database tables 42a, 42b, 42c, 42d, 42e generated from calling the sequence 41 with the proof tree 40, labeled t1 in Fig. 5, together with the example 38, labeled e1 in Fig. 5, and an initially empty set of tables. Each node of the proof tree 40 results in a database table and the path to each node is used to form the name of the database table. The example e1 was used to create the node 43 of the proof tree t1. Some examples may generate more than one proof tree. The example e1 was also used to create the node 44 of the same proof tree t1 and the path is from node 43 via a first leg 48 to the node 44. The node 46 continues from the node 44 via a first leg 45. Similarly, the example e1 was used to create the node 52 and the path is from the node 43 via a second leg 54 to the node 50 and then via a first leg 56 to the node 52.

[0034] When transforming rules into database queries, it is assumed that each rule is generated by resolving upon a clause (c1, c2 etc.) in the logical theory. The generated rule can be represented by a tree, that may here be called

partial proof tree, where the clause that is resolved upon appears in the root of the tree, and the i^{th} child of a node in the tree shows by what rule the i^{th} literal obtained from the clause at the node should be resolved upon.

[0035] Fig. 6 shows a rule 58 or rule (r1) obtained by resolving upon the first clause (c1) in the logical theory 30 shown in Fig. 2, together with a corresponding partial proof tree 60. The partial proof tree 60 has nodes 62, 64 and 66 and a first leg 68 and a second leg 70. The rule 58 is more specific or narrower than the clause c1 and includes the limitations of the clauses c2 and c4. The argument color may be replaced by red and the argument value may be replaced by face. All black cards and red numbered cards are not included in the rule 58. It may then be possible to search for examples that are covered by the rule 58. In other words, a search is conducted for examples for which proof trees can be generated using the rule 58.

[0036] It is often expensive to develop the proof trees and one important feature of the present invention is that it is only necessary to build up the proof trees once because the proof trees are saved as database tables for future use.

[0037] A database query 72 is generated in the query generator 22, shown in Fig. 1, from the partial proof tree 60, to-

gether with the label of an example, as outlined in Fig. 7. It should be noted that the function for generating table names from a sequence of a clause and child number pairs has to be identical to the one used by the sequence in Fig. 3.

- [0038] Fig. 8 shows the query 72 generated from the partial proof tree 60 shown in Fig. 6 for the example label (e1). The query 72 may be used to find out or test whether rule 58 covers a pre-existing example by matching the information of the partial proof tree 60 with suitable database tables.
- [0039] Once the proof trees of an example have been transformed into database tables, and a rule has been used to generate a database query with reference to the label of the example, the coverage check is performed by executing the query with regard to the database tables. In case no solution is found (this includes the case when the query refers to a table that does not exist), it is concluded that the rule does not cover the example. In case the solution set is non-empty, it is concluded that the rule does cover the example.
- [0040] For example, as best shown in Fig. 8, the query 72 has a FROM clause 74 that corresponds to the partial proof tree

60 and which refers to the database tables 42b and 42d, a WHERE clause 76 and AND clauses 80, 82. Both table 42b and table 42d contain tuples with example e1 and the same proof tree, as tested by the clauses 76, 80 and 82, so rule 58 covers example e1.

- [0041] In case each example has at most one proof tree, the above procedure can be simplified. The tuples that are added to the tables in the process sequence shown in Fig. 4 do only require the Example field so that the Tree field can be left out.
- [0042] Fig. 9 shows a simplified table group 78 including database tables 78a, 78b, 78c, 78d, 78e generated from calling the process sequence under this assumption with the proof tree 40 in Fig. 3 together with the example label e1 and an initially empty set of tables. In this case, there is only one proof tree per example so the table group 78 may be simplified.
- [0043] The queries that are generated by the process, as shown in Fig. 7, do not need to include the conditions that the tuples concerning a particular example also concern the same proof tree so that C can be used instead of C" when constructing the query in the process shown in Fig. 7, and hence C" does not need to be calculated in the process.

[0044] Fig. 10 shows the query 84 generated under this assumption from the partial proof tree shown in Fig. 6 for the example label e1.

[0045] Fig. 11 shows that the procedure may be extended for handling special predicates. Besides predicates that are built into the system for generating proof trees, such as $=/2$ mentioned earlier, that should not be included in the proof trees when used together with the coverage check procedure, there may also be some special predicates that cannot be excluded from the proof trees. These predicates are such that they may be resolved upon, but the clauses to use are not included in the logical theory, and their exact appearance depends on values derived from the examples. An example of such a special predicate is `split_number/1`, which is used in a logical theory 86 that includes two occurrences of this predicate 87, 89, one for each argument (Weight, Length). The purpose of this predicate is to allow values derived from the examples to act as boundaries when dividing the range for the corresponding numeric variable into two intervals so that one interval is excluded.

[0046] Fig. 12 shows a rule 88 that can be obtained in two steps from the clause in Fig. 11. The above procedure can be

extended to deal with such special predicates in the following way. Values for the special predicates are included in leaves of the partial proof trees.

- [0047] Fig. 13 shows a proof tree 90 of the example reward (5,4) given the logical theory 86 in Fig. 11. When transforming a proof tree into a set of database tables using the process sequence in Fig. 4, a table generated from a sequence that ends in a leaf that contains values for a special predicate, requires extra fields (e.g., "split_number" column) and the tuple added to the table should contain the corresponding values for these fields.
- [0048] Fig. 14 shows a group of tables 92 generated for the above example when having made this modification to the process sequence in Fig.4. When having obtained a rule by resolving upon some special predicate, the conditions on the values that are introduced during these resolution steps must also be added when using the process sequence shown in Fig. 7 to generate a database query.
- [0049] Fig. 15 shows a database query 96 obtained from the rule 88 shown in Fig. 12.
- [0050] While the present invention has been described in accordance with preferred compositions and embodiments, it is to be understood that certain substitutions and alterations

may be made thereto without departing from the spirit and scope of the following claims.